

# Detection of Inter-Spread Repeat Sequence in Genomic DNA Sequence

**Hiroo Murakami**<sup>1</sup>      **Nobuyoshi Sugaya**<sup>1</sup>      **Makihiko Sato**<sup>1,2</sup>  
hiroo@ims.u-tokyo.ac.jp      sugaya@ims.u-tokyo.ac.jp      makihiko@ims.u-tokyo.ac.jp  
**Akira Imaizumi**<sup>1,3</sup>      **Sachiyo Aburatani**<sup>1</sup>      **Katsuhisa Horimoto**<sup>1</sup>  
akima@ims.u-tokyo.ac.jp      sachiyo@ims.u-tokyo.ac.jp      khorimot@ims.u-tokyo.ac.jp

- <sup>1</sup> Laboratory of Biostatistics, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan  
<sup>2</sup> Computer Science and Engineering Centre, Fujitsu Ltd., 1-9-3 Nakase, Mihama-ku, Chiba 261-8588, Japan  
<sup>3</sup> Advanced Technology Department, Fermentation and Biotechnology Laboratories, AJINOMOTO CO., INC., 1-1 Suzuki-cho, Kawasaki-ku, Kawasaki-shi 210-8681, Japan

## Abstract

Various types of periodic patterns in nucleotide sequences are known to be very abundant in a genomic DNA sequence, and to play important biological roles such as gene expression, genome structural stabilization, and recombination. We present a new method, named “*STEPSTONE*”, to find a specific periodic pattern of repeat sequence, inter-spread repeat, in which the tandem repeats of the conserved and the not-conserved regions appear periodically. In our method, at first, the data on periods of short repeat sequences found in a target sequence are stored as a hash data, and then are selected by application of an auto-correlation test in time series analysis. Among the statistically selected sequences, the inter-spread repeats are obtained by usual alignment procedures through two steps. To test the performance of our method, we examined the inter-spread repeats in *Mycobacterium tuberculosis* and *Zamia paucijuga* genomic sequences. As a result, our method exactly detected the repeats in the two sequences, being useful for identifying systematically the inter-spread repeats in DNA sequence.

**Keywords:** tandem repeat, auto-correlation, time series analysis, inter-spread repeat

## 1 Introduction

Repeat sequences are very abundant in genomic DNA sequence, which play important biological roles, such as gene expression, genome structural stabilization, and recombination [6, 7, 8, 12, 19, 20, 21]. Tandem repeat is a type of periodic pattern and concerns several genetic diseases [3], see review [6]. In accordance with the variety of biological functions, various lengths and periodicities have been reported in the repeat sequences.

Many methods have been devised to detect different repeat sequences without *a priori* knowledge according to the characteristic features of repeat sequences. As for the repeat sequence with a short period and an exact pattern, *Tandem Repeat Finder (TRF)* [3] is one of the widely used programs, in which the probabilistic models are adopted to detect the exact repeat. To detect the long repeat sequence in a genomic scale, the *REPuter* is developed by application of the suffix-tree-based program to find a maximum size of exact repeat region [16]. In addition, the *variable number tandem repeat (VNTR)* and the *variable length tandem repeat (VLTR)* are developed to detect the repeat sequence with loose length and periodicity [14].

The repeat sequence found in the genome of *Mycobacterium tuberculosis* is an exceptional case for the patterns that the above methods are designed to detect. Indeed, the repeat sequence consists of two

regions; one is the conserved pattern sequence, and another is the random spacer sequence [10, 13]. Thus, the repeat sequence was detected not by the computational methods but by combination of the experimental studies and the visual inspection. Recently, a similar repeat sequence of conserved and non-conserved regions with a larger periodicity is also found in *Zamia paucijuga* (Cycadales) [4]. Furthermore, the knob heterochromatic structure in maize, in which the repeat sequences of conserved and not-conserved regions are detected [1], is found in the chromosome structures of *Oryza sativa* (rice) and *Arabidopsis thaliana* by experimental studies [5, 12]. Although the biological function of the repeat sequence is not still unclear, the systematic detection of the repeat is useful to facilitate the biological study.

In this study, we present a method, named *STEPSTONE* (after the fact that the conserved regions are located at regular intervals along the sequence), for detecting the periodic repeat with the conserved and not-conserved regions that is termed as inter-spread repeat. In our method, the repeat sequence is detected by focusing on the periodicity. For this purpose, at first, the data on the periods of repeat sequences are stored by the hash table for speedy analysis, and are analyzed by application of the auto-correlation in time-series analysis to the period data. Then, the repeat sequences thus selected in terms of the periodicity are further evaluated by the alignment procedures based on the sequence similarity. The performance of our method is illustrated by identifying inter-spread repeats in two genomic sequences in which the repeats are found by experimental studies and the visual inspection.

## 2 Methods

### 2.1 Definition of Repeat Sequences

In this study, two types of periodic patterns in genomic DNA sequence are focused on: the tandem repeat and the inter-spread repeat (Fig. 1). The tandem repeat is more than two periodic sequences that completely match each other. In contrast, the inter-spread repeat is more than two periodic sequences that are composed of the matching region and the not-matching region. Due to the latter region, it is difficult to detect the inter-spread repeat by usual methods for finding the periodic repeat sequences.

#### **Tandem repeat**

AGCA AGCA AGCA AGCA

#### **Inter-spread repeat**

AGCATGCC AGCAGGGT AGCAATTT AGCAAACG AGCACCCAC AGCATAGA

Figure 1: Example of tandem repeat and inter-spread repeat. The conserved regions in both repeat sequences are underlined in the sequence.

### 2.2 Algorithm of STEPSTONE

#### *Overview of Programs*

The outline of *STEPSTONE* is schematically shown in Fig. 2. At first, a target sequence is transformed into the segments of  $k$  nucleotides ( $k$ -mer). Then, the occurrence site for each  $k$ -mer on the entire region of a target sequence is recorded as a hash table, which serves to reduce the computational time in the analysis of the periodicity. Secondly, the period of  $k$ -mer is tested by a statistical method in the time series analysis. Thirdly, the repeat sequences with various periods are projected in the target sequence, and the overlapping repeats with the exactly same period are summarized in a repeat sequence. Finally, the summarized repeat sequences are aligned in consideration of evolutionary fluctuation of the not-conserved region in the inter-spread repeat. Each part of *STEPSTONE* is described below in detail.

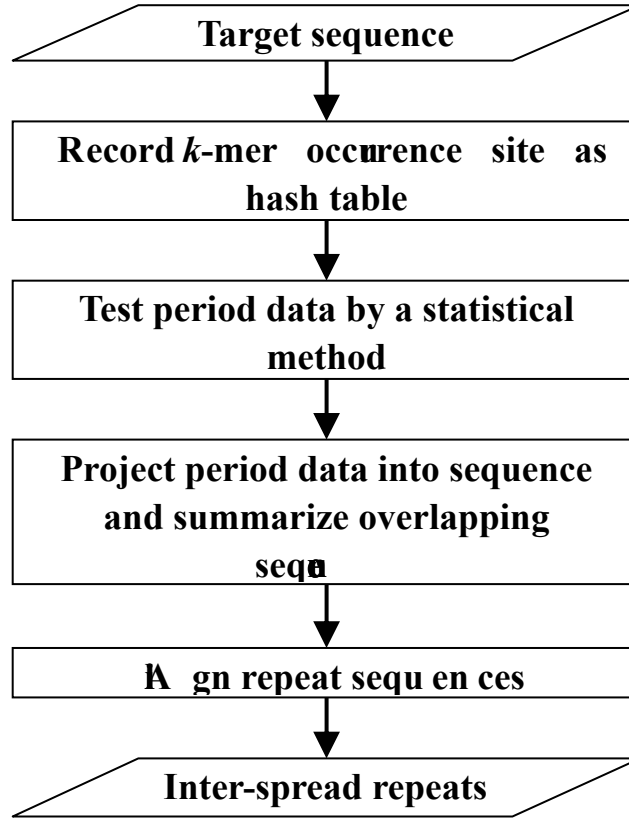


Figure 2: Flowchart of *STEPSTONE*. The *STEPSTONE* is composed of four parts enclosed rectangles.

### *Hash Table for K-Mer*

To detect repeat sequence with periodicity, a hash table of the  $k$ -mer locations is created from a target sequence and is stored in the main memory (Fig. 3). First, each nucleotide of the target sequence is converted into the numerical sequence according to a rule: from A to 0, G to 1, C to 2, and T to 3. Then, from each location of the numerical sequence, a window of  $k$  numbers is packed to the  $k$ -mer and the locations of the  $k$ -mer are stored as a hash table. This process requires only  $O(N + m)$  time:  $N$ , target sequence length;  $m$ ,  $4^k$  (number of  $k$ -mer).

### *Repeat sequence auto-correlation coefficient (RSACC)*

We test the period data of  $k$ -mers in a hash table, by application of an auto-correlation test in the time series analysis. In the application, the time in the time series analysis is regarded as the length of sequence in the repeat sequence analysis.

At first, a binary variable,  $x_n$  is defined as follows:

$$x_n = \begin{cases} 1, & \text{when the } k\text{-mer is found in the } n \text{ site of a target sequence with length } N \\ 0, & \text{otherwise} \end{cases} .$$

Then, according to the auto-correlation coefficient in time series analysis [2], we define the repeat sequence auto-correlation coefficient (RSACC),  $r_{RSACC}(l)$ , as follows:

$$r_{RSACC}(l) = \frac{E[(x_n - \bar{x})(x_{n+l} - \bar{x})]}{\sigma_x^2}, \quad (1)$$

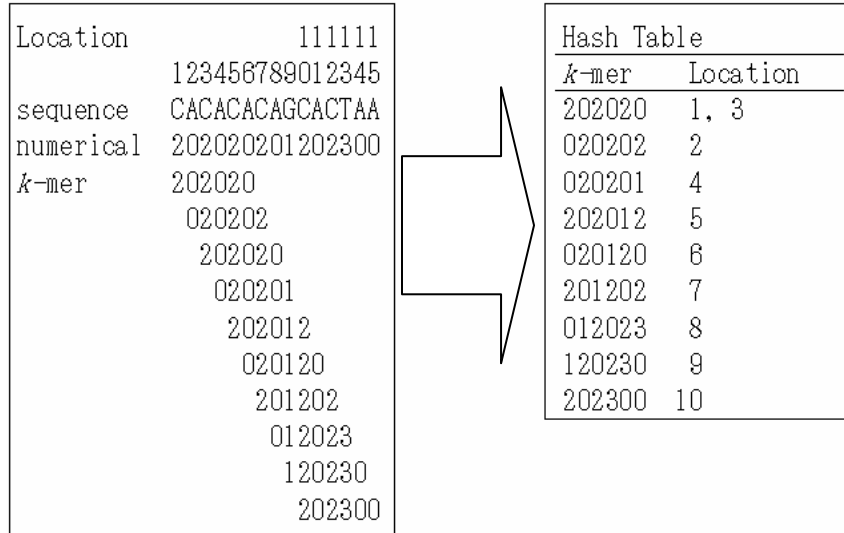


Figure 3: Example of a hash table of 6-mer in a target sequence of 15 nucleotides.

where  $\bar{x}$  and  $\sigma_x^2$  are the average and variance of  $x_n$ , i.e.,

$$\bar{x} = \frac{1}{N - k + 1} \sum_{n=1}^{N-k+1} x_n$$

and

$$\sigma_x^2 = \frac{1}{N - k + 1} \sum_{n=1}^{N-k+1} (x_n - \bar{x})^2 .$$

In the above calculation, the hash table of  $k$ -mer and its location is fully utilized. The  $\mathbf{r}_{RSACC}(l)$  in the equation (1) is tested for the null hypothesis,  $\mathbf{r}_{RSACC}(l) = 0$ , by Ljung and Box statistics [17],  $Q(l)$ , which is derived from the  $\mathbf{r}_{RSACC}(l)$  as follows:

$$Q(l) = N(N + 2) \sum_{k=1}^l \frac{\mathbf{r}_{RSACC}(k)^2}{N - k} , \tag{2}$$

and is distributed as  $\chi_l^2$ . Thus, we can test the period  $l$  of various  $k$ -mers in a target sequence.

### ***Reduction of Redundancy about Overlapping Repeat Sequence***

Among the statistically selected repeats in terms of the period, some repeats of various  $k$ -mers with the exactly same period are overlapped when the repeats are projected in a target sequence. As the next step, we reduce the redundancy about the repeats described below (Fig. 4).

The overlapped repeats in a target sequence are easily searched from the hash table. Then, among a set of overlapping repeats of different  $k$ -mers with the exactly same period, the preliminary match scores for each repeat sequence are calculated by un-gapped alignment with scoring matrix [3] in a convenient way. Among the repeat sequences thus scored, we simply select a repeat sequence that shows the maximum match score. Furthermore, some repeat sequences even with the maximum score are poorly aligned since the repeat sequence to detect in this study is much longer than initial  $k$ -mer. In this case, the alignment showing less than threshold score is filtered out. Thus, the sets of repeat sequences with the exact same period are summarized.

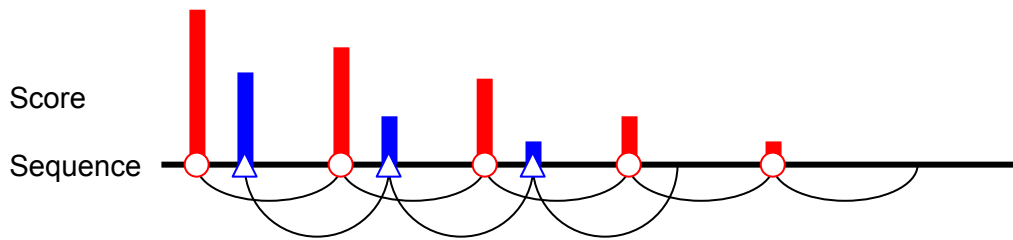


Figure 4: Selection of overlapping repeat. Two repeat sequences with the same period including distinctive  $k$ -mer (shown by a circle and a triangle) are overlapped. Each score of repeat sequences is obtained by un-gapped alignment, and the scores are shown on the  $k$ -mers by bars.

### *Refinement of Inter-Spread Repeat*

Among the non-overlapping repeat sequences with different periods, we further assemble the repeat sequences that mutually show similar periods by more than 90% and are located in the range within the less than 10% of period. The permission for the period and the location serves to consider the variation of the inter-spread repeat due to the evolutionary fluctuation especially in the not-conserved region of inter-spread repeat. Then, among each set of repeat sequences thus assembled, a repeat sequence with the maximum score obtained in the preceding step is regarded as a query sequence. The repeat sequences in each set are aligned with the query sequence by the Smith-Waterman-Goto type of dynamic programming [22].

In addition, the sets of refined alignments are classified into two types of repeat sequences: the tandem repeat and the inter-spread repeat. The relative entropy is calculated at the sites in the alignment. When the sites showing less than 50% of maximum score are successively found, the region is regarded as a not-conserved region in the inter-spread repeat. Thus, the repeat sequence is classified into the inter-spread repeat when the non-conserved region is found, and when not found, the repeat sequence is into the tandem repeat.

## 2.3 Implementation

The *STEPSTONE* algorithm is implemented by standard C++ compilers: Microsoft Visual C++ 6.0, GNU g++, and SUN developer CC with minor adjustment.

The input of a target sequence is loaded in a standard FASTA format. The parameters and their default values are as follows: length of segment ( $k$ -mer), 8; maximum length of a target sequence, 10000; DP scoring matrix, match 2, mismatch  $-5$ , gap insertion is  $-7$ , gap extension  $-5$ .

The output contains the following contents: type of repeat sequence, period and alignment with the relative entropy at each site, and the form of output is tab-separated file and/or GFF (General Features File) format file. Additionally, a graph of preliminary match score in each position of a query sequence can be output to BMP (Window's bitmap) and/or CSV (comma separated) file.

## 3 Results

We applied the *STEPSTONE* to two DNA sequences containing the inter-spread repeat sequences. In the two sequences, the inter-spread repeats were found not by the computational approaches but by experimental studies and visual inspection, in the pervious studies.

### 3.1 Inter-Spread Repeat in *Mycobacterium Tuberculosis*

In the genomic sequence of *Mycobacterium tuberculosis*, it is well known that there is a region of the inter-spread repeats (termed as *direct repeat cluster* in the original paper [13, 15]) that are composed

of the conserved sequence of 36 bp and the non-conserved spacer sequence of 34 to 41 bp. To illustrate the performance of the *STEPSTONE*, we applied it to the genomic sequence of *M. tuberculosis* H37Rv DRC region (5,189 bp, EMBL accession number: Z48304).

A part of selection process for repeat sequence in *STEPSTONE* is illustrated in Fig. 5. At first we found 270 types of 8-mers that occurred more than two times in the target sequence, with 1,398,870 periods. Then, 219 types of 8-mers with 167, 615 periods were selected among them by Ljung and Box test with 5% significance probability. In Fig. 5a, the Ljung and Box test is illustrated by two types of 8-mer. In the two 8-mers, most of 8-mers with a short range of periods and a middle range of periods were removed from the candidates of repeat sequence periods by the test, respectively. By the reduction of redundancy for the overlapping repeat sequence, 14 repeat sequences were selected from the 219 types of 8-mer in Fig. 5b. As seen in the figure, the repeat sequences with various preliminary scores are concentrated in the three regions. Indeed, 3 alignments were obtained as inter-spread repeats by the final refinement. The three alignments include the repeat sequence with almost the same periodicity: 8.95 repeats with 73 bp period, 5.96 repeats with 72 bp period, and 11.47 repeats with 73 bp period.

The projection of final alignments to the target sequence is shown in Fig. 6. As seen in the figure, three inter-spread repeat sequences detected by *STEPSTONE* (937-1639, 1934-2362, and 4018-4854 in Fig. 6a) cover about two-thirds of the regions found as inter-spread repeats in the previous study (765-1699, 1719-2462, and 3815-5153 in Fig. 6b). Note that the conserved and not-conserved regions in the detected inter-spread repeat agree well with those in the previous study. In this case, the discrimination between the conserved and not-conserved regions operates properly in the judgment of the tandem and the inter-spread repeats.

### 3.2 Inter-Spread Repeat in *Zamia paucijuga* (Cycadales)

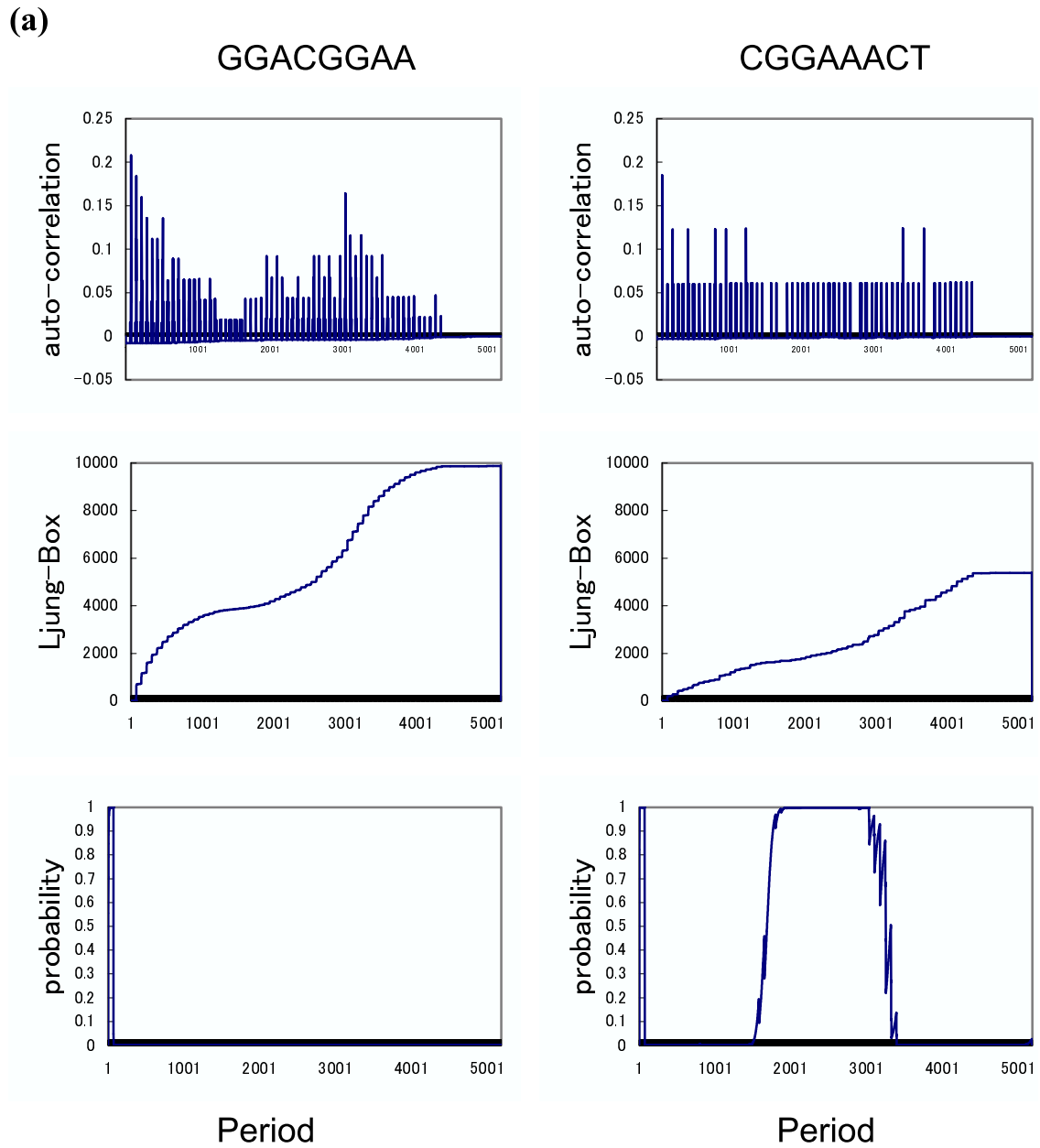
Another example is a sequence from *Zamia paucijuga*, an ancient and rare group of seed plants. We applied *STEPSTONE* to *Eco* RI fragment of ZpS1 sequence (1,563 bp, GenBank, accession number: AJ416334), in which the inter-spread repeat-like sequences of 54 bp motif and AT-rich region were found [4].

In this analysis, 335 types of 8-mer with 173, 613 periods were selected from 349 types of 8-mer with 549,840 periods found in the target sequence by Ljung and Box test with 5% significance probability. By the following reduction of overlapping repeat, 184 repeat sequences were selected. Interestingly, only one alignment was obtained by the final selection. The alignment contains a tandem repeat of periodicity 323 in Fig. 7a. The tandem repeat covers the entire region of inter-spread repeat-like sequences in Fig. 7b. Since the AT-rich regions are well conserved against the visual inspection, the *STEPSTONE* outputs the repeat sequences as the tandem repeat. At any rate, the *STEPSTONE* recognizes a longer repeat sequence that is impossible to find by eye.

## 4 Discussion

We present a systematic method, *STEPSTONE*, for detecting the inter-spread repeat of conserved and not-conserved region, for which it is difficult to be detected by previous repeat-finding methods. The validity of our method is illustrated by two sequences containing the inter-spread repeat that were detected by experiments and visual inspection. Although the performance is compared in only two examples, the *STEPSTONE* has the superior potential to *TRF* in the inter-spread repeat, and has the same potential as *TRF* in the tandem repeat with a long period in summary. The agreement of the detected and the known inter-spread repeats indicates that *STEPSTONE* is a useful tool to find a specific repeat sequence.

To compare the performance of our method to previous methods, *Tandem Repeat Finder (TRF)* [3] was applied to the two sequences analyzed in this study. In the sequence of *M. tuberculosis*, the



**(b)**

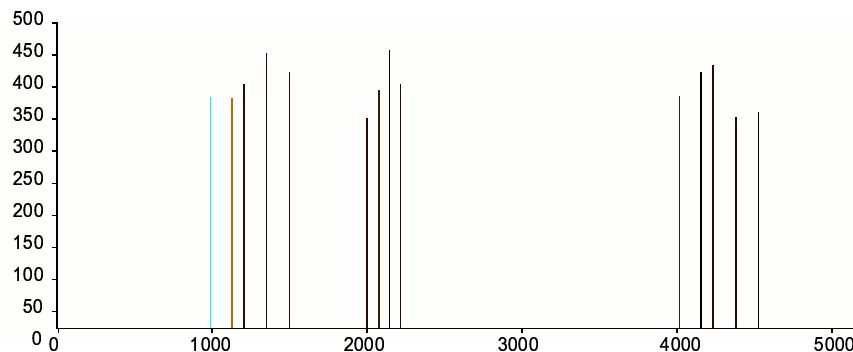


Figure 5: Selection process of inter-spread repeats in *M. tuberculosis* sequence. (a) Plots of repeat sequence auto-correlation coefficient in the equation (1), Ljung and Box statistics in the equation (2), and the probability in the  $\chi^2$  distribution along the periods in two 8-mers (left, GGACGGAA; right, CGGAAACT). (b) Preliminary scores in 14 repeat sequences. The horizontal axis indicates the target sequence, and the vertical axis indicates the score.

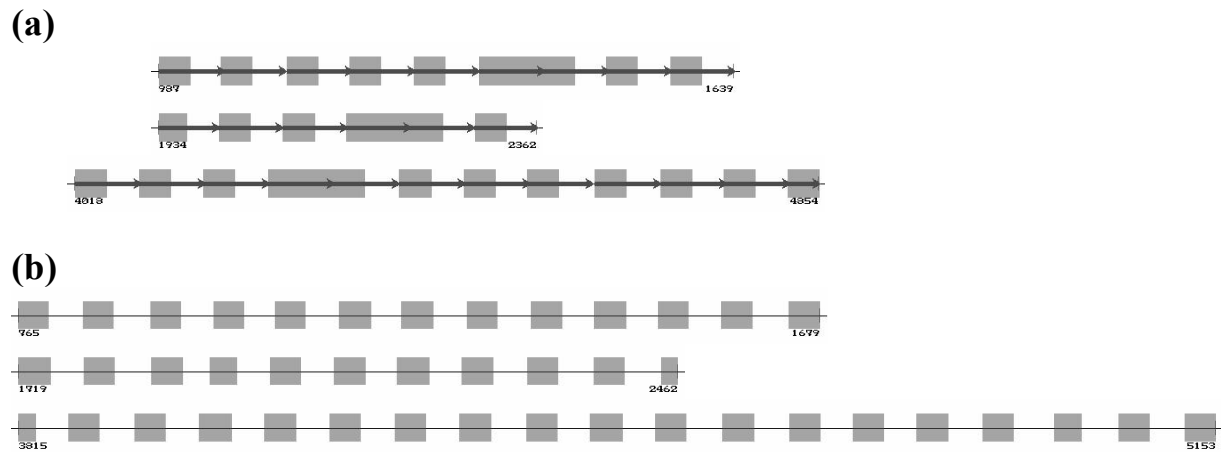


Figure 6: Correspondence between predicted and known inter-spread repeats in *M. tuberculosis* sequence. (a) Three inter-spread regions detected by *STEPSTONE* are shown. In each inter-spread repeat, a thick arrow corresponds with a repeat sequence, and a box represents a conserved region of repeat sequence in the alignment. (b) Three inter-spread repeats indicated in the previous study by experimental studies and visual inspection [13, 15]. Each box corresponds with the conserved motif.

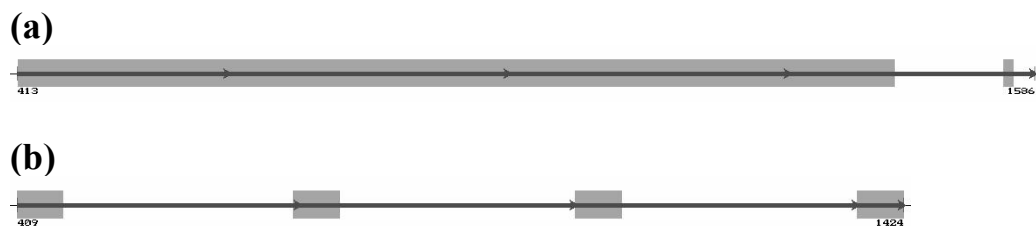


Figure 7: Correspondence between predicted and known inter-spread repeats in *Z. paucijuga* ZpS1 sequence. (a) Tandem repeat detected by *STEPSTONE*. (b) Inter-spread repeat indicated in the previous study [4]. An arrow indicates a repeat sequence, and a box indicates 54 bp long motif.

*TRF* (version 3.21 2003-01-16) with default parameters detected only one short region from 2292 to 2463 including 2.3 repeat sequences with 75 bp period (Fig. 8a). Since the *TRF* is designed to detect the tandem repeat sequences conserved over the entire region, the partial inter-spread repeats of the conserved and the not-conserved region were detected. In the *Zamia paucijuga* sequence, the *TRF* detected a similar region detected by the *STEPSTONE*: the region from 414-1424 including 3.1 repeat sequences with 323 bp period (Fig 8b). This is because the repeat sequences share the considerably conserved patterns. Although the performance is compared in only two examples, the *STEPSTONE* has the superior potential to *TRF* in the inter-spread repeat, and has the same potential as *TRF* in the tandem repeat with a long period.

To further test the performance of *STEPSTONE*, we applied it to complete sequence of *M. tuberculosis* H37Rv genome (4,411,529 bp, GenBank, accession number: NC\_000962.1). The *STEPSTONE* detected 26 inter-spread repeat sequences expect for 3 repeats in this study (data not shown). Interestingly, 25 of 26 repeats were found in the partial coding regions of PE and PPE gene family [9, 10, 11, 18]. In PE and PPE family proteins, the polymorphic multiple tandem copies of motifs are known as the possible source of antigenic variation and interfere with immune responses by inhibiting antigen processing [10]. Thus, the *STEPSTONE* may be useful to detect the repeat motifs that exist commonly in a protein family.

Our program outputs a standard GFF format, and it is easily parsed to the other annotation program. CGI version is also available in near future. The time consumed to accomplish the process

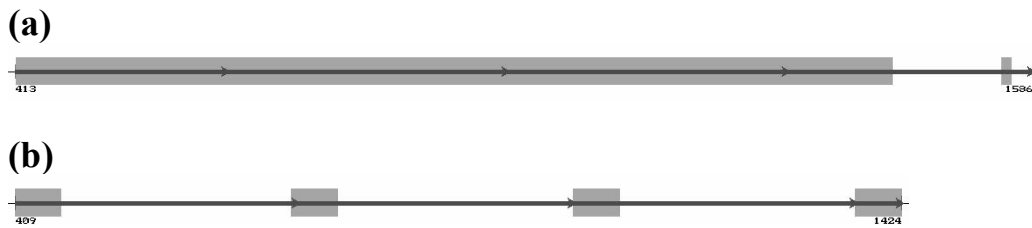


Figure 8: Schematic diagram of tandem repeats detected by *TRF*. Both an arrow and a box indicate a repeat unit. (a) A tandem repeat by applying to the sequence, *M. tuberculosis* Z48304 to *TRF*. (b) A tandem repeat by applying to the sequence, *Z. paucijuga* AJ416334 to *TRF*.

for the two sequences analyzed in this study is less than two seconds in *STEPSTONE* by using Athlon MP 1900+ PC.

## Acknowledgments

We are grateful to Drs. Natsuiro Ichinose and Tetsushi Yada (Kyoto University) for their collaborations in the early stage of this work. One of the authors (K. H.) was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas “Genome Information Science” (grant 15014208) and for Scientific Research (B) (grant 15310134), from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

## References

- [1] Ananiev, E.V., Phillips, R.L., and Rines H.W., Complex structure of knob DNA on maize chromosome 9: Retrotransposon invasion into heterochromatin, *Genetics*, 149:2025–2037, 1998.
- [2] Anderson, T.W., *The Statistical Analysis of Time Series*, New York: Wiley, 1971.
- [3] Benson, G., Tandem repeats finder: A program to analyze DNA sequences, *Nucleic Acids Res.*, 27(2):573–80, 1999.
- [4] Cafasso, D., Cozzolino, S., Luca, P.D., and Chinali, G., An unusual satellite DNA from *Zamia paucijuga* (Cycadales) characterised by two different organisations of the repetitive unit in the plant genome, *Gene*, 311:71–79, 2003.
- [5] Cheng, Z., Stupar, R.M., and Gu, M., A tandemly repeated DNA sequence is associated with both knob-like heterochromatin and a highly decondensed structure in the meiotic pachytene chromosomes of rice, *Chromosoma*, 110:24–31, 2001.
- [6] Cleary, J.D. and Pearson, C.E., The contribution of cis-elements to disease-associated repeat instability: Clinical and experimental evidence, *Cytogenet. Genome Res.*, 100:25–55, 2003.
- [7] Cohen, S., Menut, S., and Me’chali, M., Regulated formation of extrachromosomal circular DNA molecules during development of *Xenopus laevis*, *Mol. Cell. Biol.*, 19:6682–6689, 1999.
- [8] Cohen, S., Regev, A., and Lavi, S., Small polydispersed circular DNA (spcDNA) in human cells: Association with genomic instability, *Oncogene*, 14:977–985, 1997.
- [9] Cole, S.T. and Poulet, S., Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in *Mycobacterium tuberculosis*, *Arch. Microbiol.*, 163:87–95, 1995.

- [10] Cole, S.T., *et al.*, Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature*, 393:537–544, 1998.
- [11] Espitia, C., Lacleste, J.P., Mondragon-Palomino, M., Amador, A., Campuzano, J., Martens, A., Singh, M., Cicero, R., Zhang, Y., and Moreno, C., The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins?, *Microbiol.*, 145:3487–3495, 1999.
- [12] Fransz, P.F., Armstrong, S., Jong, J.H., Parnell, L.D., Drungen, C., Dean, C., Zabel, P., Bisseling, T., and Jones, G.H., Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: Structural organization of heterochromatic knob and centromere region, *Cell*, 100:367–376, 2000.
- [13] Groenen, P.M., Bunschoten, A.E., Soolingen, D., and Embden, J.D.A., Nature of DNA polymorphism in the direct repeat cluster of *Mycobacterium tuberculosis*; application for strain differentiation by a novel typing method, *Mol. Microbiol.*, 15(5):1057–1065, 1993.
- [14] Hauth, A.M. and Joseph, D.A., Beyond tandem repeats: complex pattern structures and distant regions of similarity, 18 Suppl. 1:S31–S37, *Bioinformatics*, 2002.
- [15] Hermans, P.W.M., Soolingen, D., Bik, E.M., Haas, P.E.W, Dale, J.W., and Embden, J.D.A., Insertion element IS987 from *Mycobacterium bovis* BCG is located in a hot-spot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains, *Infect. Immun.*, 59(8):2695–2705, 1991.
- [16] Kurtz, S. and Schleiermacher, C., REPuter: Fast computation of maximal repeats in complete genomes, *Bioinformatics*, 15(5):426–427, 1999.
- [17] Ljung, G.M. and Box, G.E.P., On a measure of lack of fit in time series models, *Biometrika*, 65(2):297–303, 1978.
- [18] Okkels, L.M., Brock, I., Follmann, F., Agger, E.M., Arend, S.M., Ottenhoff, T.H.M., Oftung, F., Rosenkrands, I., and Andersen, P., PPE Protein (Rv3873) from DNA segment RD1 of *Mycobacterium tuberculosis*: Strong recognition of both specific T-cell epitopes and epitopes conserved within the PPE family, *Infect. Immun.*, 71(11):6116–6123, 2003.
- [19] Ohki, R., Oishi, M., and Kiyama, R., Preference of the recombination sites involved in the formation of extrachromosomal copies of the human alphoid Sau3A repeat family, *Nucleic Acid Res.*, 23(24):4971–4977, 1995.
- [20] Pont, G., Degroote, F., and Picard, G., Some extrachromosomal circular DNAs from *Drosophila* embryos are homologous to tandemly repeated genes, *J. Mol. Biol.*, 195:447–451, 1987.
- [21] Rocha, E.P.C., Danchinb, A., and Viaria, A., Functional and evolutionary roles of long repeats in prokaryotes, *Res. Microbiol.*, 150:725–733, 1999.
- [22] Smith, T.F., Waterman, M.S., and Fitch, W.M., Comparative biosequence metrics, *J. Mol. Evol.*, 18:38–46, 1981.